Text Extraction Engine to Upgrade Clinical Decision Support System

Merve Kevser Gökgöl [1], Zeynep Orhan [2], Ajla Kirlić [3]
[1,2]Department of Information Technologies, International Burch University, Sarajevo, Bosnia and Herzegovina
[3]Department of Engineering, American University in Bosnia and Herzegovina

ABSTRACT : *New generation technological improvements lead patients to search their symptoms and corresponding diagnosis on online resources. In this study, it is aimed to develop a machine learning model to suit in different availability of users. Most of the current systems allow people to choose related symptom in web interfaces. In addition to these applications it is aimed to implement a new technique which extracts the text-based symptoms and its related parameters such as, severity, duration, location, cause, accompanied by any other indicators. This study is applicable for patient`s everyday language statements besides medical expression of symptoms for corresponding symptoms which is also supporting initial clinical decision system. Extracted terms are analyzed for matching diagnosis where an accuracy of 90% has been accomplished.*

KEYWORDS : *Medical diagnosis, symptom extraction, public healthcare, machine learning*

I. INTRODUCTION

The era of digitization has led computers to become the real face of handling commercial processes across a wealth of industries. As institutions today specifically in the medical domain have resorted to these virtual machines for realizing their goals, more and more medical data is being generated on a continuous basis (Davis et al., 2008). Current systems are challenging to optimize the utilization of existing medical data resources in automated diagnostics. The detection and interpretation of pathological conditions usually required large number of experts available, however the number of experts is sometimes not enough, and other problems may appear such as disagreement among experts (Spyns et al., 1998).

The electronic patient records contains a rich source of valuable clinical information, which could be used for a wide range of automated applications aimed at improving the health care process, such as alerting for potential medical errors, generating a patient problem list, and assessing the severity of a condition (Friedman et al., 2004). However, these applications are not applicable since large amount of information is in textual form (Tange et al., 1998).

II. METHODOLOGY

In this study, process of data collection allows patients to enter their symptoms by typing in everyday language. Therefore, to increase the accuracy firstly it is required to clean and eliminate insignificant words, such as stop words and vague abbreviations. From highest to lowest, predicted diseases are printed on the screen with matched symptoms.

Data Set : We have considered two sources for disease data. One of them is Mayo Clinic website (Mayoclinic.org., (2018)) which contains highly detailed information about diseases: disease overview, symptoms, when to see a doctor, causes, risk factors, complications and so on. The second data source is database of AZ Symptoms website (Azsymptoms.com., (2018)) where disease definitions, symptoms, causes, prevention, risk factors and complications information included for 120+ diseases. After preprocessing, each user symptom is compared with symptoms in symptom dataset. Python's Fuzzy Wuzzy library is used for similarity checking. The resulting text is saved to a new variable. These steps are both applied to the disease dataset and user text each symptom and other data valued words including severity, duration, location, cause, accompanied by any other symptoms, change in intensity are also extracted from written expression (as an individual expression or sentence structure of symptoms) accordingly. The structure of collected data categorized in four main branches; symptoms, diseases, tests (medical examinations) and corresponding treatments as described in Table1.

Symptoms	Diseases	Tests	Treatments
ID	ID	ID	ID
Name	Name	Name	Name
Nocation ID	Description	Definition	Definition
Level ID	Symptoms ID	Why_ is_ done	
Causes	Test ID	Preparation	
See_ doctor	Treatment ID	Expectation	
	Risks	Risks	
	Causes	Results	
	Prevention		
	Complications		
	Gender		
	Age		

Table 1. General structure of database table

Once tables are created symptoms are analyzed as input and they are trained by the process to detect possible diagnoses, tests and treatments. Disease data table as indicated in Table 2 consists of ID, Name, Description, Symptoms ID, Tests ID, Treatments ID, Risks, Causes, Preventions and Complications. Data content of symptoms, tests and treatments are represented with numerical values in different databases and they are embedded to system to analyze strength of the relationship.

ID	24
Name	Leukemia
Description	Leukemia is cancer of the body\'s blood-forming tissues, including the bone marrow and the lymphatic system.
Symptoms ID	21,69,114,119,133,168,173,0,0,0,0,0,0,0,0
Test ID	8,94,104
Treatment ID	113,114,115,116,117
Risks	Factors that may increase your risk of developing some types of leukemia include: Previous cancer treatment. People who\'ve had certain types of chemotherapy and radiation therapy for other cancers have an increased risk of developing certain types of leukemia. Genetic disorders. Genetic abnormalities seem to play a role in the development of leukemia. Certain genetic disorders, such as Down syndrome, are associated with an increased risk of leukemia. Exposure to certain chemicals. Exposure to certain chemicals, such as benzene â€” which is found in gasoline and is used by the chemical industry â€” also is linked to an increased risk of some kinds of leukemia. Smoking. Smoking cigarettes increases the risk of acute myelogenous leukemia. Family history of leukemia. If members of your family have been diagnosed with leukemia, your risk for the disease may be increased'
Causes	Scientists don`t understand the exact causes of leukemia. It seems to develop from a combination of genetic and environmental factors.
Prevention	None
Complications	None
Gender	Not gender restricted
Age	Not age restricted

Table 2. Diseases table in database

Symptoms are identified by ID value, name, locations (neck, knee, eye, low back), level of the symptom which is also indicated with numerical values (sharp, low, high, sudden, mild) and causes, and possible conditions for a patient to visit the doctor. Table of tests (medical examinations) includes details about essential examinations` ID, name, definition, why it is required (why is done), preparation, expectation, risks, results. On the other, hand table of recommended treatments for the most related treatments are classified with as their ID numbers, names and definitions.

Implementations : A smart text extracting clinical decision support engine is developed to convert the clinical data into significant and effective information. Python is used to develop the most efficient and appropriate model. TextBlob is a library used for input and output processing, and for string matching which actually classifies input symptoms as a disease. Two sets are used for classification, one including only symptoms, and the other the matching diseases. Details about diseases such as treatments and tests are recommended, are stored in MySQL database. The user is asked to enter his/her symptoms, and then the application converts the answer into Text Blob object. Fuzzy String Matching, also called approximate string matching, is implemented as the process of finding strings which approximately match a given pattern. The closeness of a match is often measured in terms of edit distance which is the number of primitive operations necessary to convert the string into an exact match.

Preprocessing : Each symptom information is sent to the preprocessing module to prepare the data for the analysis. First, stop words, punctuation characters and blank spaces are removed from the text. Then, each word is stemmed by Porter's Stemmer. The resulting text is saved to a new variable. These steps are both applied to the disease dataset and user text

Training and testing : In training process text-based symptoms, patient`s personal information and past medical history is used as input data. In the next step, data is trained for the possible detected diagnoses and recommend appropriate treatment and as a result of training we expect to get out model output.

After the training process, the system is tested for various diagnoses and patients, then the percentage risk of possible disease is represented as output information. Corresponding to this, treatments and recommendations (any medical examinations, tests) are driven.

III. RESULTS AND CONCLUSION

A text extracting clinical decision support engine is developed to increase the accuracy of medical diagnosis. The project may provide a significant help in clinical decision process which gives effective results even with patients own words and in their own language. The behavior of different classifiers was tested in the context of the problem. classifying inputted symptoms into predefined disease classes. The study was approached with three different methods: using naïve Bayes, decision tree and Fuzzy Wuzzy library. The goal was to make a system that will give as correct classification as possible regardless of spelling mistakes. Different inputs were tested to assess the abilities supported by the Text Blob library. Output is based on the result obtained using Fuzzy Wuzzy library regardless of some spelling mistakes that user might have done in giving input. After testing the system, an accuracy of 90% has been accomplished. The impact on outcomes, assessing whether the project reduces time from diagnosis to treatment, reduces cost, and improves quality the benefits of the study in global health care environment.

Future Study Objectives: In process of developing the engine more detailed information about the patient will also be collected to evaluate the model efficiently and give more accurate prediction of treatments and recommendations. These are patients past medical history (allergies, medicines, surgeries, family history, diets, and habits), birth and growth information, age, gender, height, weight. In order to improve the current system, we also aim to upgrade the engine available without restriction of input language.

REFERENCES

1. Davis, D. a., Chawla, N. V., Blumm, N., Christakis, N., Barabasi, A.-L., & Barabási, A. (2008). Predicting individual disease risk based on medical history. Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM '08, 769. https://doi.org/10.1145/1458082.1458185
2. Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association, 11(5), 392–402. https://doi.org/10.1197/jamia.M1552
3. Singh, H., & Sittig, D. F. (2015). Setting the record straight on measuring diagnostic errors. Reply to: "Bad assumptions on primary care diagnostic errors" by Dr Richard Young. BMJ Quality & Safety, 24(5), 345.2-348. https://doi.org/10.1136/bmjqs-2015-004140
4. Spyns, P., Nhàn, N. T., Baert, E., Sager, N., & De Moor, G. (1998). Medical language processing applied to extract clinical information from dutch medical documents. Studies in Health Technology and Informatics. https://doi.org/10.3233/978-1-60750-896-0-685
5. Tange, H. J., Schouten, H. C., Kester, A. D., & Hasman, A. (1998). The granularity of medical narratives and its effect on the speed and completeness of information retrieval. Journal of the American Medical Informatics Association : JAMIA, 5(6), 571–82. https://doi.org/10.1136/jamia.1998.0050571

6. Steven Loria (2017). TextBlob: Simplified text processing [a Python (2 and 3) library for processing textual data]. Retrieved from http://textblob.readthedocs.io/en/latest/index.html
7. Mayoclinic.org. (2018). Diseases and Conditions - Disease and condition information from Mayo Clinic experts. [online] Available at: https://www.mayoclinic.org/diseases-conditions/index [Accessed 24 Feb. 2018].
8. Azsymptoms.com. (2018). Medical Symptoms: Online Encyclopedia of Medical Symptoms and Conditions. [online] Available at: http://www.azsymptoms.com/ [Accessed 24 Feb. 2018].